nature genetics

Population structure, differential bias and genomic control in a large-scale, case-control association study

David G Clayton[1], Neil M Walker[1], Deborah J Smyth[1], Rebecca Pask[1], Jason D Cooper[1], Lisa M Maier[1], Luc J Smink[1], Alex C Lam[1], Nigel R Ovington[1], Helen E Stevens[1], Sarah Nutland[1], Joanna M M Howson[1], Malek Faham[2], Martin Moorhead[2], Hywel B Jones[2], Matthew Falkowski[2], Paul Hardenbol[2], Thomas D Willis[2] & John A Todd[1]

The main problems in drawing causal inferences from epidemiological case-control studies are confounding by unmeasured extraneous factors, selection bias and differential misclassification of exposure[1]. In genetics the first of these, in the form of population structure, has dominated recent debate[2–4]. Population structure explained part of the significant +11.2% inflation of test statistics we observed in an analysis of 6,322 nonsynonymous SNPs in 816 cases of type 1 diabetes and 877 population-based controls from Great Britain. The remainder of the inflation resulted from differential bias in genotype scoring between case and control DNA samples, which originated from two laboratories, causing false-positive associations. To avoid excluding SNPs and losing valuable information, we extended the genomic control method[2–5] by applying a variable downweighting to each SNP.

The results reported here concern the first phase of a study that will eventually include 8,000 cases of type 1 diabetes (T1D; Genetic Resource Investigating Diabetes study)[6,7] and 8,000 controls drawn from the 1958 British Birth Cohort. The aim of this first phase was to create a short list of the nonsynonymous SNPs (nsSNPs) most closely associated with disease for testing in larger numbers of subjects in later phases; on its own, this phase is underpowered to detect all but the strongest associations (odds ratios > 1.7), which we believe will be rare in common, multifactorial diseases[8]. This project was an analysis of more than 12,000 nsSNPs in 2,000 DNA samples using the newly described, highly multiplexed MegAllele technology[9,10]. Given the magnitude of the project and our expectation that we would be searching for effect sizes mainly in the range of odds ratios of 1.1–1.5 (ref. 8), which could easily be generated or confounded artifactually by subtle and hidden biases, we wanted to limit potential biases that could arise from shifts in any of its components, including subtle variations in liquid handling, reagents, software and technology. DNA samples, which were all from lymphoblastoid cell lines, were extracted using the same protocol in two different laboratories. Case and control DNAs were arranged randomly, and the teams were blinded to the case-control status so that case-control status could not have an effect on sample processing during genotyping. Initially, we scored genotypes of cases and controls together, as blind scoring is normally regarded as a prerequisite for inclusion of studies in meta-analyses. We tried to obtain as much data as possible from as many of the nsSNP assays as possible: each nsSNP lost is a potential causal variant for T1D lost to the study.

Our previous extensive experience with manually scored, singleplex SNP technologies, Taqman and Invader[6,7,11], indicated that target sequences with SNPs in the genome vary widely in their ease of allelic discrimination and PCR efficiency, thereby producing a wide range in the extent of clustering within the clouds of fluorescent signal data points for the three genotype classes. We routinely reject singleplex assays that show overlap between either of the homozygous clouds and the heterozygous one (called half-calls) upon visual inspection; this is not a cost-efficient approach when thousands of SNPs are assayed simultaneously. Here, we scored genotypes automatically. Our initial results showed some extreme associations that could not be replicated with singleplex genotyping (data not shown). We observed that, for some nsSNPs, the positions of the data clouds differed between cases and controls (**Fig. 1a**,**b**), resulting in spurious case-control differences. In extreme cases, instead of three clouds representing the two homozygous and one heterozygous genotypes, we observed four clouds: the heterozygous case and control clouds were quite distinct (*e.g.*, the *CD44* SNP rs9666607; **Fig. 1a**). When separation of the signal clouds corresponding to different genotypes is anything other than perfect, automated scoring algorithms face a dilemma: attempting to score a high proportion of samples runs the risk of assigning erroneous genotypes. But adopting the conservative strategy of calling only genotypes with a high degree of certainty (perfectly separated clouds) is also not free from problems. Failure to call will not be independent of genotype, a phenomenon called ‘informative missingness’, leading to bias in allele and genotype frequencies that are estimated using only called samples. When, as here, signal clouds for cases and controls are displaced, the likely result with either strategy is differential bias

[1]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge, CB2 2XY, UK. [2]ParAllele BioScience, 7300 Shoreline Court, South San Francisco, California 94080, USA. Correspondence should be addressed to D.G.C. (david.clayton@cimr.cam.ac.uk) or J.A.T. (john.todd@cimr.cam.ac.uk).

Received 10 June; accepted 15 August; published online 9 October 2005; doi:10.1038/ng1653

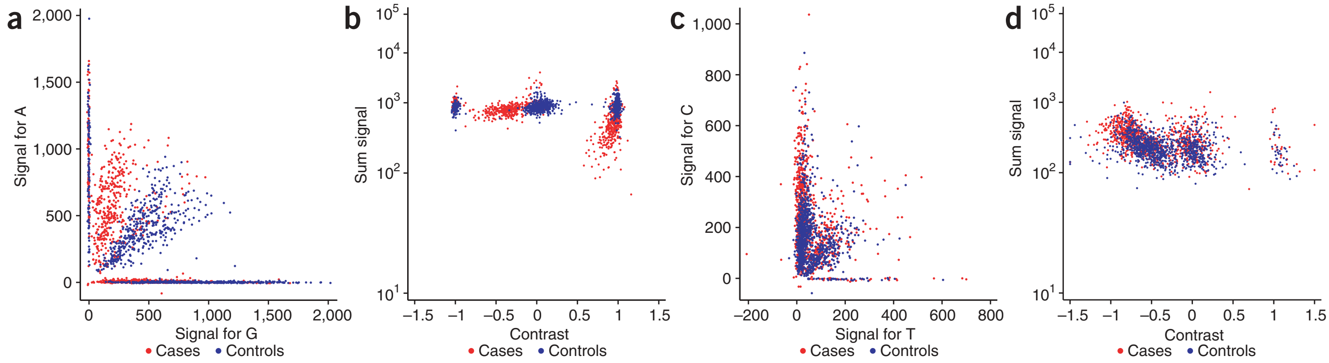

Figure 1 Signal intensity plots. (**a**) Signal intensity plot for *CD44* SNP rs9666607. (**b**) Transformed signal intensity plot for *CD44* SNP rs9666607. (**c**) Signal intensity plot for *IL13* SNP rs2054. (**d**) Transformed signal intensity plot for *IL13* SNP rs2054.

resulting in false-positive association (*e.g.*, the *IL13* SNP rs2054[1]; **Fig. 1c,d**). Even SNPs that show 100% genotype concordance between two different genotyping platforms could suffer from differential bias. Moreover, despite the fact that all DNA samples came from cell lines and were obtained using the same extraction protocol, underlying, unexplained differences remained between the two sample sets in the interaction of DNA source and the allelic discrimination chemistry. Therefore, we modified our strategy by scoring cases and controls separately to adapt to cloud shifts. For calibration of typing rules, the use of a single, standard set of DNA samples (such as a HapMap set[12]) would not have solved this difficulty and could have exacerbated it.

We generated a quantile-quantile plot of observed Cochran-Armitage test statistics for disease association versus each of the 6,322 common autosomal nsSNPs scored in 816 cases and 877 controls (**Fig. 2**). The 168 nsSNPs in the HLA region were excluded, because many of them showed the expected very strong associations with T1D. We also excluded any results with extreme deviation from Hardy-Weinberg equilibrium in either the controls or the cases, because this could be indicative of a missing genotype cloud due to serious genotyping failure. The tests are ranked from smallest to largest and plotted against their expected values under the global null hypothesis that no true association exists. There were still some extreme results, which was unexpected given the small sample size, and evidence of general inflation of the test statistics, as shown in the plot by the line of slope $\lambda = 1.112$ (+11.2%; 95% confidence interval = 1.069–1.157) obtained by fitting a line to the lower 90% of the distribution (**Fig. 2**). Scoring cases and controls separately did not fully solve the problem of differential bias due to displacement of the genotype clouds and might even have itself contributed to overdispersion of test statistics. We studied the determinants of this overdispersion using generalized linear modeling and found that it was best predicted by a combination of the half-call rate and the difference between call rates for cases and controls. By restricting the analysis to the 4,620 nsSNPs with half-call rates of <0.25% and difference in call rates within ±5%, the inflation of association test statistics was reduced to +5.0% (95% confidence interval = 1.003–1.098; **Fig. 3**).

Population structure could be a cause of the remaining 5% inflation in this subset of 4,629 nsSNPs[2–4]. Although both case and control samples were drawn from throughout Great Britain, there were differences in their geographical distribution, with cases outnumbering controls by almost 2:1 in Scotland and in northern England, and controls outnumbering cases by a similar margin in London and in southern and southeastern England. Stratified tests in which cases and controls are compared within geographical regions showed minimal overdispersion (+1.5%; 95% confidence interval = 0.969–1.062) for the 4,629 nsSNPs (**Fig. 4**). Once between-region variation in nsSNP allele frequencies in the British population was taken into account (by matching cases and controls within 12 regions of Great Britain), there was little or no evidence of residual population structure effects in test statistic inflation.

The differential bias affected many nsSNPs to varying degrees. Imposing rigorous quality thresholds, by removing 1,693 nsSNPs with imperfect genotype clouds from the analysis, risks discarding potentially valuable data and missing true positive findings that would not be included in phase 2 genotyping (in a second, independent set of cases and controls). To emphasize this point, imposing the even more

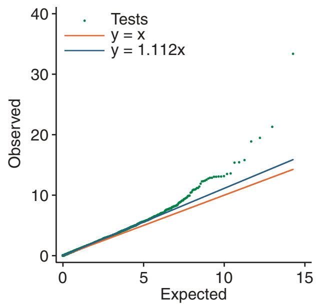

Figure 2 Quantile-quantile plots of Cochran-Armitage test statistics. The ranked, observed values for 6,322 nsSNPs are plotted against the values expected for sampling from a χ^2 distribution with one degree of freedom (the distribution expected under the null hypothesis).

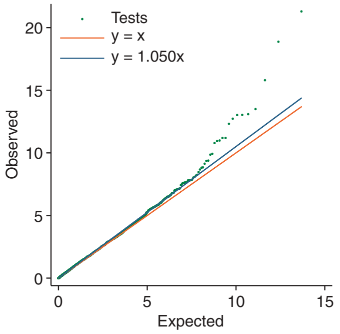

Figure 3 Quantile-quantile plots of Cochran-Armitage test statistics of 4,629 nsSNPs with half-call rates <0.5% and a difference in call rates between cases and controls of no more than 5%.

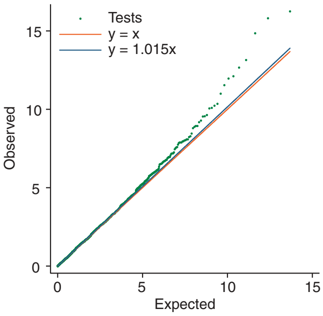

Figure 4 Quantile-quantile plots of tests for association for 4,629 nsSNPs after stratification by broad geographical region.

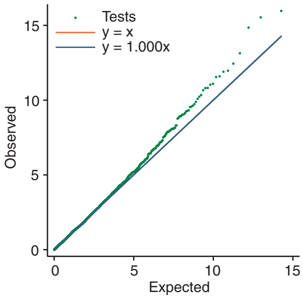

Figure 5 Quantile-quantile plots of geographically stratified test statistics for 6,322 nsSNPs with downweighting as a function of half-call rate and differential call rate.

stringent limit of 0.1% for half-call rates left 2,943 nsSNPs for which the overdispersion was only +0.4%, but with the loss of considerable data. A compromise is suggested by the idea of genomic control for sub structure[5] in which χ^2 statistics are downweighted by multiplying them by a factor $1/\lambda$, where λ is the variance inflation factor. This idea may be extended to deal with differential genotyping calling: the generalized linear model that we used to explore the predictors of overdispersion of test statistics can also generate a weighting factor for each nsSNP (a function of half-call rate and absolute difference in call rates) so that overdispersion is reduced to the level observed for the nsSNPs with lowest half-call and highest call rates (**Supplementary Note** online). This procedure allowed us to use the data from all 6,322 nsSNPs ($\lambda = 1.000$; **Fig. 5**). The weighting factors varied in magnitude from 0.46 for the 79 nsSNPs with the lowest quality clouds to 1.0 for the 3,271 best nsSNPs; most nsSNPs (94.4%) had weighting factors of >0.8. There is a loss of power for the $\sim$5% of nsSNPs that were downweighted, but this is preferable to excluding them completely. It remains possible that the model for overdispersion could be improved, perhaps by including metrics derived from fluorescence intensity measures. This would further improve the specificity of downweighting.

We have no reason to believe that the phenomenon that led to differential bias in our study is specific to this genotyping platform. The best way to avoid increased false-positive rates due to such biases is to control all aspects of DNA source, preparation and genotyping using the paradigms of blindness and randomization, but this will not always be possible. The success of the selective downweighting approach described here depends on the availability of informative diagnostics from automated scoring software, as applying a uniform, overall correction to critical values to control the type 1 error rate would result in unnecessary loss of power for reliably scored SNPs. For example, the effect of the observed differential bias was to inflate χ^2 tests by $\sim$6%; extrapolating this to our final 8,000 cases and 8,000 controls would suggest at least 50% inflation. Use of a single correction to control the type 1 error to 10^{-6} on a perfectly typed marker would reduce a 90% power to less than 60%, a 70% power to 30%, and a 50% power to less than 20%. Therefore, it is preferable to apply such corrections more selectively. There is a need for better diagnostics for genotyping scoring software and, perhaps, for more flexibility in allowing discrimination rules to adapt to sample characteristics. We also showed that power can be preserved in British case-control studies by matching cases and controls by specific geographical regions.

The highly multiplexed genotyping platforms to be used in whole-genome association studies will inevitably be subject to errors in assigning genotypes. Exclusion of ambiguous calls is no remedy, as this too can distort recorded genotype frequencies. Coupled with subtle differences in sample preparation, which may be impossible to exclude, these biases can lead to increased false-positive rates. These may outweigh the effects of population substructure, which have long been a concern for population-based studies. As for the effects of substructure, the ideal solution is to exclude such problems by using improved methods. Although some assays may be so poor that they must be excluded, rather less serious and more widespread biases due to differential misclassification and differential call rate may remain. The idea of genomic control can then be extended to correct the consequent effects on the false-positive rate.

METHODS

Subjects. The case sample comprised 1,024 cases of T1D from the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory's Genetic Resource Investigating Diabetes study. This is a subsample of a resource that is projected to reach 8,000 by the end of 2006. Most cases (64.6%) were <16 years of age at the time of collection of the samples, all resided in Great Britain, and all were of (self-reported) European descent. The control sample comprised 1,023 participants in a cohort study of all births in Great Britain during a single week in 1958. The control sample donors resided throughout Great Britain and >97% were of European descent.

DNA preparation. DNA samples were processed in two different laboratories using the same protocol. Blood peripheral lymphocytes were isolated from whole blood and used to create immortalized lymphoblastoid lines using Epstein Barr virus. These lines were then expanded to produce cell pellets for DNA extraction. The DNA was extracted from these cell pellets using an organic extraction method, which uses chloroform to isolate the DNA and ethanol to precipitate it. Resultant DNA was resuspended in Tris-EDTA buffer and quantified using Picogreen Reagent. On receipt of the control DNA at the Diabetes and Inflammation Laboratory, it was requantified using the same Picogreen method used to quantify the case DNA to supply the assay with equal amounts of both case and control DNA. After requantification, samples were normalized to a concentration of 150 ng μl^{-1} as required by the Affymetrix genotyping platform and interdigitated into 96-well plates.

Genotyping reactions. Molecular inversion probes were designed and ordered from ParAllele Bioscience. The genotyping reactions were carried out using the standard protocol recommended by the manufacturer and as described[9,10], at Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory and at ParAllele Bioscience. Molecular inversion assays were carried out in 96-well plates using 24 individuals per plate for each of four allele channels using the MegAllele genotyping kit (ParAllele BioScience). Comparison and follow-up genotyping was carried out using the singleplex Taqman MGB chemistry (ABI).

Automated genotype calling. Genotypes were scored at ParAllele Bioscience using Euclidean clustering analysis of the 'contrast' measures derived from the normalized signal intensities using the EM algorithm. For a SNP with alleles A and G and signal intensities I_A and I_G, respectively, sum signal = $I_A + I_G$ and contrast $= \frac{\sinh\left(\frac{2(I_A I_G)}{I_A + I_G}\right)}{\sinh(2)}$. In this analysis, for each SNP, estimates of the genotype frequencies and of the mean and variance of the signal contrast measure for each genotype are computed. At the same time, 'posterior' probabilities of possible individual genotype assignments are calculated. Positive calls are made if the ratio of posterior probabilities for the most probable call and the next most probable call exceeds a threshold. A half-call is made when this condition is not met but when one of the homozygous genotypes can be excluded at the same threshold. Although half-calls are classified as missing genotypes, the half-call rate is a useful parameter in finding clusters susceptible to genotyping bias.

As genotype clusters for cases and controls were displaced for some markers, we scored case and control genotypes separately. But this could be expected to introduce some inflation of test statistics. As outlined above, individual calls are guided to some extent by estimates of genotype frequencies. By scoring separately, we allowed these estimates to differ between cases and controls; the effect of this is to inflate the variance of case-control contrasts, in much the same way as "cryptic relatedness" introduces correlation between cases and between controls, but not between cases and controls[5]. This effect will be most pronounced for those SNPs in which clusters are overlapping, as calls are then more dependent on estimated genotype frequencies; this explains why the half-call rate is a good predictor of overdispersion.

Selection of SNPs. In May 2004 we identified 17,599 nsSNPs by a search of dbSNP version 120; 12,340 were attempted and 10,469 passed a fluorescence signal threshold in a test plate containing local controls and 55 DNA samples from the HapMap[12] Centre d'Etude du Polymorphisme Humain panel. Concordance with Illumina BeadArray technology was 99.2% for the 2,980 nsSNPs in common in HapMap release 8. The scoring software incorporated four quality control criteria: call rate $>$80%; half-call rate $<$10%; signal to background ratio $>$30; and controls coefficient of variation $<$30%. Any case or control sample not satisfying all of these criteria failed. For 816 cases and 877 controls, 9,025 nsSNPs were judged to have working assays according to the above minimum quality control metrics, and 8,134 of these were polymorphic. We excluded 168 nsSNPs in the HLA region (25–35 Mb on chromosome 6p) and deferred analysis of 45 nsSNPs on the X chromosome. Of the remaining 7,931 nsSNPs, 6,539 had minor allele frequency $>$1%. We instituted further quality control checks to detect gross failures, such as nondetection of one or more genotype clusters. Another ten nsSNPs were rejected because they were monomorphic in cases or controls, and an additional 207 were rejected owing to gross departure from Hardy-Weinberg equilibrium in one or both groups ($\chi^2 > 16$). Of the Hardy-Weinberg failures, 118 were consistent with missing genotype clusters; of the remaining 89, 31 failed in both cases and controls, 31 in controls only and 27 in cases only. It is possible, though unlikely, that one or more of the extreme HWE failures in the cases is a genuine disease association, and we intend to investigate this possibility.

Statistical methods. All statistical tests used were asymptotically distributed as χ^2 with one degree of freedom under the null hypothesis. Quantile-quantile plots were produced by plotting the ranked values of the test statistic against the (approximate) expected order statistic, $F^{-1}[i \,/\, (N + 1)]$, where $F()$ is the $\chi^2(1)$ cumulative distribution function and $i = 1\ldots N$ represent ranks from smallest ($i = 1$) to largest ($i = N$) value. We estimated an overdispersion factor, λ, by calculating the ratio of the mean of the smallest 90% of observed test statistics to the mean of the corresponding expected values. Test statistics plotted in **Figures 2** and **3** are Cochran-Armitage one-degree-of-freedom tests for the 3 × 2 tables formed by cross-classifying subjects by genotype and disease status[13]. To allow for geographic variation of the case:control ratio (**Figs. 4** and **5**), we carried out stratified comparison within 12 geographic regions, using the one–degree-of-freedom stratified trend test proposed[14]. We studied the determinants of overdispersion using a generalized linear model[15], specifying a gamma family for the distribution of test statistics and logarithmic link for the effects of the predictors of overdispersion. We tried many possible predictors, including call rates in both cases and controls, average signal strengths and half-call rates (signals that lay between the main homozygous and heterozygous clusters). Two independently significant predictors were found: the half-call rate and the absolute difference in call rates between cases and controls. To obtain the corrected values plotted in **Figure 5**, half-call rates were classified as a four-level factor (0, up to 0.25%, 0.25–5%, or $>$5%) and the difference in call rates was classified as a binary factor ($\leq$5% or $>$5%). This model was highly predictive of overdispersion (χ^2 due to model = 53.58 on 6 degrees of freedom). We then used the fitted model to calculate a multiplicative correction factor for each group of tests to correct their mean to that for the 'perfect' category (0 half-call rate and differential call rate $<$0.25%). Additional information about the genotype calling procedures can be found in **Supplementary Note** and ref. 16.

URLs. Information about the Genetic Resource Investigating Diabetes study is available at http://www-gene.cimr.cam.ac.uk/ucdr/grid.shtml. Information about the 1958 British Birth Cohort is available at http://www.cls.ioe.ac.uk/. dbSNP version 120 is available at http://www.ncbi.nlm.nih.gov/projects/SNP/. Genotype data are available through the Juvenile Diabetes Research Foundation/ Wellcome Trust Diabetes and Inflammation Laboratory homepage at http://www-gene.cimr.cam.ac.uk/todd/. Results were visualized using the T1D-specific genome browser, T1DBase (http://www.t1dbase.org/).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank the individuals with T1D and control individuals for their participation; G. Coleman, S. Field, T. Mistry, K. Bourget, S. Clayton, M. Hardy, P. Lauder, M. Maisuria, W. Meadows and S. Wood for preparing DNA samples; D. Strachan, R. Jones, S. Ring and W. McArdle for providing DNA from the 1958 British Birth Cohort collection; and A. Long, N. Naclerio, T. Cormier, K. Tran, C. Bruckner and S. Picton for genotyping and technical assistance. We acknowledge use of DNA from the 1958 British Birth Cohort collection, funded by the Medical Research Council and the Wellcome Trust. We thank the Juvenile Diabetes Research Foundation, the Wellcome Trust, Diabetes UK and the Medical Research Council for financial support. D.G.C. is a Juvenile Diabetes Research Foundation and Wellcome Trust Principal Research Fellow.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Breslow, N.E. & Day, N.E. *Statistical Methods in Cancer Research* Vol. I. *The Analysis of Case-Control Studies* (International Agency for Research on Cancer, Lyon, 1980).
2. Devlin, B., Bacanu, S.A. & Roeder, K. Genomic control to the extreme. *Nat. Genet.* **36**, 1129–1130; author reply 1131 (2004).
3. Freedman, M.L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).
4. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
5. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
6. Vella, A. *et al.* Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76**, 773–779 (2005).
7. Lowe, C.E. *et al.* Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun.* **5**, 301–305 (2004).
8. Wang, W.Y., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118 (2005).
9. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
10. Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**, 269–275 (2005).
11. Ueda, H. *et al.* Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
12. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
13. Armitage, P. Test for linear trend in proportions and frequencies. *Biometrics* **II**, 375–386 (1955).
14. Mantel, N. Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.* **58**, 690–700 (1963).
15. Nelder, J. & Wedderburn, R. Generalised linear models. *J. R. Statist. Soc. A* **135**, 370–384 (1972).
16. Moorhead, M. *et al.* Optimal genotype determination in highly multiplexed SNP data. *Eur. J. Hum. Genet.* (in the press).